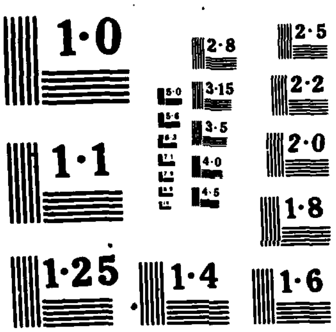
AD-A168 349   ON THE USE OF RIDGE AND STEIN-TYPE ESTIMATORS IN   1/1
PREDICTION(U) STANFORD UNIV CA DEPT OF STATISTICS
A E GELFAND 21 MAY 86 TR-374 N00014-86-K-0156

UNCLASSIFIED   F/G 12/1   NL



NATIONAL BURFAU OF S
MICROCOPY RESOLU1 TEST

12

ON THE USE OF RIDGE AND STEIN-TYPE ESTIMATORS IN PREDICTION

BY

ALAN E. GELFAND

TECHNICAL REPORT NO. 374
MAY 21, 1986

PREPARED UNDER CONTRACT
N00014-86-K-0156 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

ON THE USE OF RIDGE AND STEIN-TYPE ESTIMATORS IN PREDICTION

BY

ALAN E. GELFAND

TECHNICAL REPORT NO. 374

MAY 21, 1986

Prepared Under Contract
N00014-86-K-0156 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

ON THE USE OF RIDGE AND STEIN-TYPE ESTIMATORS IN PREDICTION

Alan E. Gelfand

1. Introduction

For the usual regression model with fixed regressors, $Y = X\beta + \varepsilon$, $Y_{n\times 1}$, $X_{n\times p}$ full rank, $\beta_{p\times 1}$ and $\varepsilon_{n\times 1} \sim (0,\sigma^2 I)$, there is considerable literature devoted to alternatives to the ordinary least squares estimator, $\hat{\beta}_{OLS}$ of β. From work originally dating to Stein (1956) and James and Stein (1961) when ε is normally distributed and $p \geq 3$, $\hat{\beta}_{OLS}$ is inadmissible under loss $(\hat{\beta}-\beta)^T Q(\hat{\beta}-\beta)$, Q an unrestricted positive definite matrix. Thus, much of this extant discussion focuses on the development of biased estimators with small "variances" which achieve a smaller expected loss *either uniformly over* p-dimensional Euclidean space or at least in the vicinity of some specified β^*. Two "classes" of such reduced variance regression estimators are particularly well discussed - ridge estimators and Stein-type estimators. Either directly or upon orthogonal transformation these estimators take the form

$$(1) \qquad \hat{\beta}_C = C\hat{\beta}_{OLS} + (I-C)\beta^*$$

where C is a diagonal matrix, usually data dependent. They may also be seen to be Bayes or "Empirical" Bayes procedures as well. The review paper by Draper and Van Nostrand (1979) provides an excellent summary of both the theoretical and simulated effort in this area.. In the context of cross-validation, i.e. of examining the performance of an estimator obtained in one sample in prediction in a second independent sample, the work of Stone (1974) leads to estimators of the form in (1) as well.

Herein we consider the simplest such cross-validation problem. At a new vector of predictor values, X_0, we seek to estimate $X_0^T\beta$. We take as loss function $(\delta(Y)-X_0^T\beta)^2$ for an estimator $\delta(Y)$ and we assume henceforth that ε is normally distributed with σ^2 unknown. Our problem differs from that of estimating the vector β since the results of Cohen (1965) show that $\alpha X_0^T\hat{\beta}_{OLS}$ is an admissible estimator of $X_0^T\beta$ for $0 \leq \alpha \leq 1$, i.e. the UMVU estimator is admissible. (In fact, $\delta(Y)$ of the form $\gamma^T Y$ is admissible for $X_0^T\beta$ i.f.f. $(2\gamma-X(X^TX)^{-1}X_0)^T(2\gamma-X(X^TX)^{-1}X_0) \leq X_0^T(X^TX)^{-1}X_0$.) Nonetheless, if we have some confidence in β^*, i.e. that β^* is near the true value β, then it makes sense to attempt to improve upon $X_0^T\hat{\beta}_{OLS}$ in the "vicinity of β^*" using estimators of the form (1). More specifically, how well do the "classes" of ridge estimators and of Stein-type estimators perform in this prediction? Can we make a "best" choice within these classes for a particular prediction?

The problem of prediction of an independent observation Y_0 at X_0 using the loss $(\delta(Y)-Y_0)^2$ is equivalent to that of predicting $X_0^T\beta$, i.e. $E_\beta(\delta(Y)-Y_0)^2 = \sigma^2 + E_\beta(\delta(Y)-X_0^T\beta)^2$.

For an estimator of the form $X_0^T\hat{\beta}$, the expected loss becomes

$$(2) \qquad E_\beta(\hat{\beta}-\beta)^T X_0 X_0^T(\hat{\beta}-\beta) = E_\beta[\Sigma X_{0i}(\hat{\beta}_i-\beta_i)]^2 .$$

In the sequel we take the generalized ridge estimator $\hat{\beta}_R$ to be

$$(3) \qquad \hat{\beta}_R = (X^TX+A)^{-1}(X^TY+A\beta^*)$$

where A is p.d. symmetric and possibly dependent on Y. We take the general Stein-type estimator $\hat{\beta}_S$ to be

$$(4) \qquad \hat{\beta}_S = (1- c/Q)\hat{\beta}_{OLS} + c/Q\ \beta^*$$

where $Q = (\hat{\beta}_{OLS}-\beta^*)^T X^T X(\hat{\beta}_{OLS}-\beta^*)$ and c may depend on Y. In practice c/Q is usually replaced by min(c/Q, 1).

In section 2 we calculate the risk, (2), of the estimators (3) and (4) when A, c are constant. We then investigate "best" choices for A, c. Since these choices will be functions of β and σ^2 as well as X_0, A and c must be estimated from Y. In section 3 we summarize a simulation study which compares the performance of versions of (3) and (4) which are discussed for the estimation of β along with others motivated by work in section 2. In section 4 we offer concluding remarks in particular with regard to multiple prediction.

2. Theoretical Results

We first note that for $\hat{\beta}_{OLS}$ (2) becomes

$$\sigma^2 X_0^T(X^TX)^{-1}X_0 \ . \tag{5}$$

We now claim that

Theorem 1: For $\hat{\beta}_R$ as in (3), (2) becomes

$$\begin{aligned}&\sigma^2 X_0^T(X^TX+A)^{-1}X^TX(X^TX+A)^{-1}\\ &\quad + X_0^T(X^TX+A)^{-1}A(\beta-\beta^*)(\beta-\beta^*)^TA(X^TX\ +A)^{-1}X \ .\end{aligned} \tag{6}$$

Proof: We transform to principal components form. Let R be nonsingular such that $RX^TXR^T = I$, $RAR^T = D$, D a diagonal matrix with diagonal entries d_i. For any point β in p-dimensional Euclidean space, let $\alpha = (R^{-1})^T\beta$. Then

$$\hat{\alpha}_R = (R^{-1})^T\hat{\beta}_R = (I+D)^{-1}\hat{\alpha}_{OLS} + (I+D)^{-1}D\alpha^* \ ,$$

i.e. of the form (1) with $C = (I+D)^{-1}$. In terms of α, (2) becomes $E_\alpha(\hat{\alpha}-\alpha)w_0w_0^T(\hat{\alpha}-\alpha)$, $w_0 = RX$. Since $\hat{\alpha} \sim N(\alpha,\sigma^2 I)$, this expectation is readily calculated to be

$$\sigma^2 w_0^TC^2w_0 + w_0^T(I-C)(\alpha-\alpha^*)(\alpha-\alpha^*)^T(I-C)w_0 \ .$$

Substitution for α, w_0 and C yields (6). □

Note: Normality is not employed in this calculation.

In (3) A.is usually taken to be diagonal and, in fact, the class of ridge (as opposed to generalized ridge) estimators sets $A = aI$, $a \geq 0$. The case where either by design or transformation $X^TX = I$ reduces (5) to $\sigma^2 \Sigma X_{0i}^2$ and reduces (6), for generalized ridge estimators (a_i are the diagonal elements of the diagonal matrix A) to

$$(7) \qquad \sigma^2 \Sigma X_{0i}^2 \frac{1}{(1+a_i)^2} + [\Sigma X_{0i}(\beta_i - \beta_i{}^*) \frac{a_i}{1+a_i}]^2 .$$

Investigation of this expression reveals that an optimal choice for the a_i to minimize (7) needn't exist although local minima can be found. In the case of ridge estimation, i.e. all $a_i = a$, a unique minimum can be found. This occurs at

$$(8) \qquad a_0 = \sigma^2 \gamma^{-2} X_0^T X_0$$

where $\gamma = X_0^T(\beta - \beta^*)$. Note that $a_0 > 0$ and finite provided $\beta - \beta^*$ isn't orthogonal to X_0. The associated minimum equals

$$\frac{\sigma^2 \gamma^2 X_0^T X_0}{\gamma^2 + \sigma^2 X_0^T X_0} .$$

When β is such that $\beta - \beta^*$ is orthogonal to X_0, then $X_0^T\beta^*$ predicts perfectly. For such β's we can obtain zero expected loss and would want no weight attached to $\hat{\beta}_{OLS}$, i.e. would want $a = \infty$. In fact, it is clear that for X_0, β fixed there will be a set of β' 's which predict X_0^T perfectly and that β' needn't be close to β in Euclidean distance. Thus the appropriate pseudometric for the prediction problem is $(\beta_1-\beta_2)^T X_0 X_0^T (\beta_1-\beta_2)$. This pseudometric clarifies the earlier notion of "vicinity of β^*" and under this distance the further β^* is from β the closer a is to 0, i.e. the more weight is placed on $\hat{\beta}_{OLS}$, the closer β^* is to β the larger a becomes, i.e. the more weight is placed on β^*. As would be expected, a_0 is invariant to scaling of X_0, although the risk clearly isn't.

Using (8) our estimator of $X_0^T\beta$ is

$$T_{a_0} = (1+a_0)^{-1} X_0^T \hat{\beta}_{OLS} + (1+a_0)^{-1} a_0 X_0^T \beta^*$$

and, in fact, for any fixed $a > 0$, T_a improves upon $X_0^T\hat{\beta}_{OLS}$ whenever $\gamma^2 < \sigma^2 a^{-1}(2+a)$.

From (8) a convenient estimator of a_0 is:

$$\hat{a}_0 = \hat{\sigma}^2 \hat{\gamma}^{-2} X_0^T X_0 \tag{9}$$

when σ^2 is the usual UMVU estimator of σ^2 and $\hat{\gamma} = X_0^T(\hat{\beta}_{OLS}-\beta^*)$. The fact that $E\hat{\gamma}^{-2}$ doesn't exist suggests that $\hat{a}_0$ will be very unstable and that $T_{\hat{a}_0}$ will perform poorly. We return to this point in the discussion of the simulation study. Since $\hat{\sigma}^2$ is independent of $\hat{\beta}_{OLS}$ and $\hat{a}_0$ depends on $\hat{\beta}_{OLS}$ only through $X_0^T\hat{\beta}_{OLS}$, we may compute the expected loss for $T_{\hat{a}_0}$. If $\tau^2 = \sigma^2 X_0^T X_0$, then $\hat{\gamma} \sim N(\gamma,\tau^2)$ and, with $\hat{\tau}^2 = \hat{\sigma}^2 X_0^T X_0$, (2) for $T_{\hat{a}_0}$ becomes

$$(10) \quad E_\gamma\left(\frac{\hat{\gamma}^2}{\hat{\gamma}^2+\hat{\tau}^2}\,\hat{\gamma}-\gamma\right)^2 = \tau^2+E_\gamma\left(\frac{\hat{\gamma}^2(\hat{\tau}^2)^2+2\tau^2\hat{\tau}^2\hat{\gamma}^2-2\tau^2(\hat{\tau}^2)^2}{(\hat{\gamma}^2+\hat{\tau}^2)^2}\right) .$$

The equality (10) is seen using the identity $E_\gamma f(\hat{\gamma})(\hat{\gamma}-\gamma) = \tau^2 E_\gamma f'(\hat{\gamma})$ (Stein (1973)) valid provided $E_\gamma|f'(\hat{\gamma})| < \infty$ which, as the following calculations show, is the case). Now $\hat{\gamma}^2/\tau^2 \sim X_1^2$, $\gamma^2/2\tau^2$ independent of $\hat{\tau}^2/\tau^2 \sim X_{n-p}^2$. Hence $(\hat{\gamma}^2+\hat{\tau}^2)^{-1}\hat{\tau}^2|L \sim Be(\frac{n-p}{2}, \frac{2L+1}{2})$ where $L \sim Po(\gamma^2/2\tau^2)$. The expectation of each term in (10) can thus be evaluated and (10) becomes

$$(11) \quad \tau^2\{L+(n-p)E(n-p+2L+3)^{-1}[\frac{(n-p+2)(2L+1)}{(n-p+2L+5)} - \frac{n-p-2L+1}{n-p+2L+1}]\}.$$

If we divide (11) by τ^2, i.e. consider the risk relative to that of $X_0^T\hat{\beta}_{OLS}$, then this relative risk is a function of γ^2/τ^2. Hence we set $\tau^2 = 1$ and examine the simpler estimator $(\hat{\gamma}^2+1)^{-1}\hat{\gamma}^3$ which may be thought of as an "empirical" Bayes estimator against a normal prior centered at 0, adjusted to have no singulariti in R^1. The risk of this estimator is readily obtained to be $1 + E_\gamma(\hat{\gamma}^2+1)^{-2}(3\hat{\gamma}^2-2)$ by an argument similar to that leading to (10). This risk (symmetric about 0) is graphed in Figure 1 against $\gamma > 0$ to illustrate what may be expected, up to scaling, if (11) is evaluated. Note that the risk is bounded and considerably less than 1 for γ small. Because $(\hat{\gamma}^2+1)^{-2}\hat{\gamma}^3$ has singularities in the complex plane it is not admissible.

If we restore X^TX, not necessarily diagonal, our estimator in (3) has $A = a_0X^TX$ or $\hat{a}_0X^TX$ according to (8) or (9).

Theorem 2: For $\hat{\beta}_S$ as in (4) with $p > 2$, (2) becomes

$$(12)\qquad \sigma^2x_0^T(X^TX)^{-1}x_0 + x_0^T(X^TX)^{-1}x_0[(c^2+4c\sigma^2)\Gamma_1/\sigma^2-2c\Gamma_2]$$

where

$$\Gamma_1 = E\,\frac{2L+1}{(p+2(L+M))(p+2(L+M)-2)}\ ,\quad \Gamma_2 = E\,\frac{1}{p+2(L+M)-2}$$

with

$$(13)\qquad \begin{aligned} &L \sim Po(\lambda)\ ,\quad \lambda = \gamma^2/2\sigma^2x_0^T(X^TX)^{-1}x_0 \\ &M \sim Po(\delta)\ ,\quad \delta = (\Delta x_0^T(X^TX)^{-1}x_0-\gamma^2)/2\sigma^2x_0^T(X^TX)^{-1}x_0 \end{aligned}$$

where L,M independent and $\Delta = (\beta-\beta^*)^TX^TX(\beta-\beta^*)$.

Proof: As in Theorem 1, let R be nonsingular such that $RX^TXR^T = I$ and let $\hat{\alpha} = (R^{-1})^T(\hat{\beta}_{OLS}-\beta^*)$. Then $\hat{\alpha} \sim N(\alpha,\sigma^2I)$ with $\alpha = (R^{-1})^T(\beta-\beta^*)$, $Q = \hat{\alpha}^T\hat{\alpha}$, and (2) becomes

$$(14) \qquad E_\alpha[(1 - \frac{c}{Q})w^T\hat{\alpha} - w^T\alpha]^2$$

where $w = RX_0$ (and $w^Tw = X_0^T(X^TX)^{-1}X_0$).

If we expand (14) we obtain

$$(15) \qquad E_\alpha(w^T(\hat{\alpha}-\alpha))^2+c^2E(w^T\hat{\alpha})^2/(\hat{\alpha}^T\hat{\alpha})^2-2cE_\alpha\frac{(w^T\hat{\alpha})}{\hat{\alpha}^T\hat{\alpha}}\, w^T(\hat{\alpha}-\alpha)\ .$$

The last term may be written as $-2c\Sigma w_iE_\alpha f(\hat{\alpha})(\hat{\alpha}_i-\alpha_i)$ where $f(\hat{\alpha}) = (\hat{\alpha}^T\hat{\alpha})^{-1}(w^T\hat{\alpha})$. Using the Stein identity, $(\sigma^2E_\alpha\frac{\partial f(\hat{\alpha})}{\partial\hat{\alpha}_i} = E_\alpha f(\hat{\alpha})(\hat{\alpha}_i-\alpha_i)$, which is valid here), on this expression, after manipulation (15) becomes

$$(16) \qquad \sigma^2w^Tw + (c^2+4c\sigma^2)E_\alpha(w^T\hat{\alpha})^2/(\hat{\alpha}^T\hat{\alpha})^2 - 2c\sigma^2w^TwE_\alpha(1/\hat{\alpha}^T\hat{\alpha}).$$

Finally if we let $U = \frac{(w^T\hat{\alpha})^2}{w^Tw}$, $V = \hat{\alpha}^T\hat{\alpha} - \frac{(w^T\hat{\alpha})^2}{w^Tw}$, then

$\frac{U}{\sigma^2}|L \sim \chi^2_{1+2L}$ with L as in (13)

$\frac{V}{\sigma^2}|M \sim \chi^2_{p-1+2M}$ with M as in (13)

and given L and M, $\frac{U}{U+V} \sim Be(\frac{1+2L}{2}, \frac{p-1+2M}{2})$ independent of $\frac{U+V}{\sigma^2} \sim \chi^2_{p+2(L+M)}$. Hence

$$E_\alpha \frac{(w^T\hat{\alpha})^2}{(\hat{\alpha}^T\hat{\alpha})^2} = w^T w E_\alpha E(\frac{U}{(U+V)^2} \,|L,M) = w^T w E_\alpha [E(\frac{U}{U+V}|L,M)E(\frac{1}{U+V} \,|L,M)]$$

$$= \frac{w^T w}{\sigma^2} \; E \; \frac{2L+1}{(p+2(L+M))} \; \frac{1}{p+2(L+M)-2} \doteq \frac{w^T w}{\sigma^2} \Gamma_1$$

and similarly

$$E_\alpha (\frac{1}{\hat{\alpha}^T\hat{\alpha}}) = \frac{1}{\sigma^2} \Gamma_2 \; .$$

Making these substitutions into (16) and restoring X_0 we obtain (12). ▯

Note: The proof reveals that the expected loss, (2), for more general estimators of the form $(1-h(Q))\hat{\beta}_{OLS}+h(Q)\beta^*$ can be developed. In fact, if

$$E_\beta \left| \frac{\partial h(Q)\hat{\gamma}}{\partial \hat{\beta}_{OLS,i}} \right| < \infty$$

the loss is

(17) $$\sigma^2 X_0^T(X^TX)^{-1}X_0 + E_\beta h^2(Q)\hat{\gamma}^2 - 2\sigma^2 X_0^T(X^TX)^{-1}X_0 E_\beta h(Q) - 4\sigma^2 E_\beta h'(Q)\hat{\gamma}^2$$

Inspection of (12) reveals that the unique best c is

(18) $$c_0 = \sigma^2(\frac{\Gamma_2}{\Gamma_1} - 2) \; .$$

Since λ,δ are invariant to scaling of X_0 so is c_0. It is apparent that for X_0 fixed as $\Delta \to 0$, $\Gamma_2/\Gamma_1 \to p$, i.e. $c_0 \to \sigma^2(p-2)$.. Hence if we believe β^* is close to β the "usual" constant, $\sigma^2(p-2)$, may be employed. Using this constant, if Δ is near 0, the relative risk of $X_0^T\hat{\beta}_S$ to $X_0^T\hat{\beta}_{OLS}$ will, from (12), be near 2/p, as it is in the case of estimating β. As in remarks after (10), if $\hat{\sigma}^2$ is the usual estimator of σ^2 independent of $\hat{\beta}_{OLS}$ we may compute (2) for $\hat{\beta}_S$ as in (4) with $c = \hat{\sigma}^2(p-2)$.[1] We obtain

$$(19) \quad \sigma^2 X_0^T(X^TX)^{-1}X_0\{1+\Gamma_1(p^2-4+2(n-p)^{-1}(p-2)^2)-2\Gamma_2(p-2)\}.$$

Expressions similar to (19) can be obtained for instances of the more general estimators mentioned above (17). This suggests that the risk (2) may be calculated for commonly used (adaptive) ridge estimators, e.g. those considered in section 3. However, without restrictions on the design matrix X, these estimators often fail to either provide $\hat{a}$ in closed form or to define $\hat{a}$ as a function of Q.

Since to a first order approximation $c_0 \simeq \sigma^2(2\lambda+1)^{-1}(p-2+2(\delta-\lambda))$ we may estimate c_0 by

$$(20) \qquad \hat{c}_0 = \hat{\sigma}^2(2\hat{\lambda}+1)^{-1}(p-2+2(\hat{\delta}-\hat{\lambda}))$$

where $\hat{\lambda},\hat{\delta}$ are the expressions in (13) with β replaced by $\hat{\beta}_{OLS}$. We would truncate c_0 to the interval $[0,Q]$. Since $E\hat{\lambda}^{-1}$ doesn't exist $\hat{c}_0$ will be very unstable (as with $\hat{a}_0$ in (9)) suggesting that the resulting predictor will perform poorly. Again we return to this point in the next section.

3. A Simulation Study

A simulation was conducted to compare the use of the OLS predictor with the predictors discussed in the previous section and with predictors arising from other estimators of β which have been discussed in the literature. For convenience we set $\sigma^2 = 1$ and take $X^TX = I$, i.e. $\hat{\beta}_{OLS} \sim N(\beta,I)$.[2] Without loss of generality we set $\beta^* = 0$ and $X_0^TX_0 = 1$. Under this setup ridge estimators become $(1+\hat{a})^{-1}\hat{\beta}_{OLS}$ and Stein estimators become $(1 - c/\hat{\beta}_{OLS}^T\hat{\beta}_{OLS})\ \hat{\beta}_{OLS}$. In addition to $\hat{\beta}_{OLS}$ we consider the following six estimators of β (4 ridge type, 2 Stein type).

(i) $\hat{\beta}_{HK}$ - arising from $\hat{a} = p/\hat{\beta}_{OLS}^T\hat{\beta}_{OLS}$. The ridge estimators discussed by Hoerl and Kennard (1970), Hoerl, Kennard and Baldwin (1975) and, in fact, Lawless and Wang (1976) reduce to $\hat{\beta}_{HK}$ in our setup.

(ii) $\hat{\beta}_{RM}$ - arising from $\hat{a} = p/(\hat{\beta}_{OLS}^T\hat{\beta}_{OLS}-p)$ with $(1+\hat{a})^{-1} = 0$ if $\hat{\beta}_{OLS}^T\hat{\beta}_{OLS} \leq p$. The RIDGM and STEINM estimators discussed by Dempster, Shatzoff, and Wermuth (1977) reduce to $\hat{\beta}_{RM}$.

(iii) $\hat{\beta}_{MG}$ - arising from $\hat{a} = (1 - p/\hat{\beta}_{OLS}^T\hat{\beta}_{OLS})^{-1/2}(1-(1- p/\hat{\beta}_{OLS}^T\hat{\beta}_{OLS}$ with $(1+\hat{a})^{-1} = 0$ if $\hat{\beta}_{OLS}^T\hat{\beta}_{OLS} \leq p$. The ridge estimator of McDonald and Galarneau (1975) reduces to $\hat{\beta}_{MG}$.

(iv) $\hat{\beta}_{\hat{a}_0}$ - arising from $\hat{a}_0$ given in (9).

(v) $\hat{\beta}_{p-2}$ - arising from $c = p-2$, i.e. the "usual" Stein estimator.

(vi) $\hat{\beta}_{\hat{c}_0}$ - arising from $\hat{c}_0$ given in (20), truncated to $[0,\infty)$.

"Positive part" restrictions were applied to all "shrinkages" in (v) and (vi).

We note that under the above assumptions the risks in (3) and (4) and, in fact, of the predictors arising from (i) - (vi) depend on X_0 and β only through $(X_0^T\beta)^2$ and $\beta^T\beta$. Since $(X_0^T\beta)^2 = r\beta^T\beta$, $0 \leq r \leq 1$, we may summarize the results in terms of $\beta^T\beta$ and r. We consider p = 3,6,10. For a given p we generated sets of 2p independent uniform random variables on the interval [-1,1]. In each case we considered the first p observations as a p vector, standardized to length 1 and designated it as an X_0. Similarly the second p observations are considered as a β vector with scaling by .1,1,10. Hence we have large $\beta^T\beta$, i.e. $\beta^T\beta = 100$, moderate $\beta^T\beta$, i.e. $\beta^T\beta = 1$ and small $\beta^T\beta$, i.e. $\beta^T\beta = .01$. For each X_0,β pair 1000 $\hat{\beta}_{OLS}$'s were generated from $N(\beta,I)$ and using X_0 each of the seven predictors were calculated for each of the 1000 replications. Bias, variance and mean square error (MSE) were estimated. A large number of X_0,β pairs (approximately 400) were investigated enabling a wide range of r's. Table 1 provides a brief summary indicating the best $\hat{\beta}$ for prediction over ranges for r along with the typical percentage reduction in risk (using the best predictor over that range), $100(\text{MSE } X_0^T\hat{\beta}_{OLS} - \text{MSE } X_0^T\hat{\beta})/\text{MSE } X_0^T\hat{\beta}_{OLS}$.

Several comments are appropriate:

(i) The cross-over points in Table 1 are approximate, but in the vicinity of the cross-over competing predictors are indistinguishable with respect to MSE.

(ii) It is not surprising that regardless of β, if $\beta^T\beta$ large and r large, the OLS predictor is best. In fact, if $\beta^T\beta$ large and $.01 < r < .5$, the percent improvement of the best predictor over OLS is never greater than 5%.

(iii) As expected, $\hat{\beta}_{\hat{c}_0}$, $\hat{\beta}_{\hat{a}_0}$ performed very badly, always sixth or seventh, doing well only when $\beta^T\beta$ large and r very small (regardless of p). However, in such cases, improvements will be substantial, increasing as r decreases, while the other five predictors are indistinguishable. Near $r = .0$ $\hat{\beta}_{\hat{a}_0}$ is best; much below .01, $\hat{\beta}_{\hat{c}_0}$ is best.

(iv) $\hat{\beta}_{p-2}$ is likely the best overall choice always amongst the two or three best apart from cases in (iii) above.

(v) When $\beta^T\beta$ is small or moderate, $\hat{\beta}_{RM}$, $\hat{\beta}_{HK}$, $\hat{\beta}_{p-2}$ and $\hat{\beta}_{MG}$ were always the best four. When $\beta^T\beta$ is small and $p = 3$, $\hat{\beta}_{MG}$ is close in performance to $\hat{\beta}_{p-2}$. When $\beta^T\beta$ is small and $p = 6$, $\hat{\beta}_{RM}$ and $\hat{\beta}_{MG}$ split for second best. When $\beta^T\beta$ is small and $p = 10$, $\hat{\beta}_{p-2}$ is second with $\hat{\beta}_{MG}$ third.

(vi) When r is large $\hat{c}_0$ is almost always <0 whence $\hat{\beta}_{\hat{c}_0} = \hat{\beta}_{OLS}$.

Table 1 reveals that in many cases substantial reduction in squared error loss over the OLS predictor can be achieved. It further suggests the possibility of selecting the predictor according to $\beta^T\beta$ and r. However, finely detailed selection, e.g. according to X_0, will be unsuccessful as the performance of $\hat{\beta}_{\hat{c}_0}$ and $\hat{\beta}_{\hat{a}_0}$ reveals. In practice we will have Δ/σ^2 instead of β^T and $r = (\Delta X_0^T(X^TX)^{-1}X_0)^{-1}Y^2$, and we might define estimators $\hat{\Delta},\hat{r}$ with β replaced by $\hat{\beta}_{OLS}$, σ^2 replaced by $\hat{\sigma}^2$. We may calculate $E(\hat{\Delta}) = \frac{n-1}{n-3}(p+\Delta)$ whence $\hat{\Delta}' = \frac{n-3}{n-1}\hat{\Delta}-p$ is UMVU for Δ. By an argument similar to that contained in Theorem 2, we may show $E(\hat{r}) = E(p+2(L+M))^{-1}(2L+1) \simeq r+(1-rp)\{(p+\Delta)^{-1}+(p+\Delta)^{-3}2\Delta\}$ where L,M are distributed as in (13). For individual predictions, preliminary calculation of $\hat{\Delta}'$ and $\hat{r}$ should enable a judicious choice of predictor.

As Thisted and Morris (1980, p. 19) observe, the poorest estimation case for ridge procedures occurs when (with $X^TX = I$, $\beta^* = 0$) $\beta^T = (\beta_1,0,0,\ldots,0)$ with β_1 large. This is also the poorest estimation case for Stein type procedures in the sense that the first coordinate will account for about half of the total risk and all coordinate risks would decrease if $\hat{\beta}_1$ was excluded (see Baranchik (1964)). For prediction this implies Δ large and $r = X_{01}^2/X_0^TX_0$. Hence this is the poorest case for prediction as well, i.e. depending upon X_{01}, improvement will be small or, in fact, the OLS predictor will be better.

How will multicollinearity in X^TX affect prediction? Let $X^TX = D$, a diagonal matrix with diagonal elements d_i and assume $d_1 \leq d_2 \leq \ldots \leq d_p$. Then the extent of multicollinearity is usually measured in terms of how close d_1 is to 0. Although ridge methods have been advocated for improved estimation when d_1 is quite small, particularly relative to the other d_i, Thisted and Morris (p. 21) and others have established that in this case optimal ridge as well as Stein-type estimators will produce inconsequential improvement over $\hat{\beta}_{OLS}$. For prediction (with $\beta^* = 0$), $\Delta = \Sigma d_i\beta_i^2$ and $r = (X_0^T\beta)^2/(\Sigma X_{0i}^2/d_i \cdot \Sigma d_i\beta_i^2)$. Bingham and Larntz (1977, p. 102) observe that (in this notation) the worst case for ridge estimation occurs when large β_i are associated with small d_i. This tells us little about the magnitude of Δ. However for fixed X_0 and β as X^TX becomes more severely multicollinear $\mathrm{var}(X_0^T\hat{\beta}_{OLS}) = \Sigma X_{0i}^2/d_i$ will grow larger and r will become smaller. As the simulation suggests when r is small, using an appropriate predictor, we can expect significant improvement over the OLS predictor.

4. Multiple Prediction

In concluding we offer several comments regarding multiple or simultaneous prediction. Suppose we wish to make r predictions defined by $X_1, X_2, \ldots, X_r$ and we set $X^* = (X_1, \ldots, X_r)$. For convenience we assume the X_i are a linearly independent set whence rank $(X^*) = r \leq p$. What is an appropriate loss to employ? One choice is unweighted sum of squared error loss, i.e. $\Sigma(X_i^T(\hat{\beta}-\beta))^2 = .(\hat{\beta}-\beta)^T G_1(\hat{\beta}-\beta)$ with $G_1 = X^*X^{*T}$. A second choice arises from the joint distribution of the $X_i^T\hat{\beta}_{OLS}$, i.e. $X^{*T}\hat{\beta}_{OLS} \sim N(X^{*T}\beta, \sigma^2(X^{*T}(X^TX)^{-1}X^*)^{-1})$ suggests $(\hat{\beta}-\beta)^T G_2(\hat{\beta}-\beta)$ where $G_2 = X^*(X^{*T}(X^TX)^{-1}X^*)^{-1}X^{*T}$. Others may be envisioned as well. If $r < p$, G_1, G_2 are positive semi-definite whence, as noted earlier for individual prediction, we may have loss equal to 0 but $\hat{\beta}$ not close to β. Nonetheless, it is well-known that if P is any positive definite matrix $X^{*T}\hat{\beta}_{OLS}$ is admissible for $X^{*T}\beta$ under loss $(\hat{\beta}-\beta)^T X^*PX^{*T}(\hat{\beta}-\beta)$ if $p \leq 2$, inadmissible if $p \geq 3$. In fact, work done by Berger (1976), Bhattacharya (1966), Bock (1975), Casella (1977), Efron and Morris (1976) and Strawderman (1978) leads to explicit minimax predictors which improve upon $X^{*T}\hat{\beta}_{OLS}$. These predictors will be generalized adaptive ridge of the form $(I + A(X^{*T}\hat{\beta}_{OLS}, \hat{\sigma}^2, X^{*T}(X^TX)^{-1}X^*, P))^{-1}X^{*T}\hat{\beta}_{OLS}$, (see e.g. Strawderman, Theorem 6, p. 626, for a family of such predictors), parelleling, in a sense, the individual predictors $X_0^T\hat{\beta}_{\hat{a}_0}$ and $X_0^T\hat{\beta}_{\hat{c}_0}$.

As Strawderman notes (p. 626) there is no one predictor which will dominate the OLS predictor for all P. For $P = I$ (i.e. G_1) a very simple procedure is to use estimates of Δ and r to select a good predictor for each X_i. If we are not prepared to specify P the simulation study suggests that using $\hat{\beta}_{p-2}$ (i.e. $c = \hat{\sigma}^2(p-2)$ in (4)) regardless of X_i is a simple but perhaps adequate choice.

References

Baranchik, A. (1964). "Multiple Regression and Estimation of the Mean of Multivariate Normal Distribution," Dept. of Statistics, Stanford University, Technical Report #51.

Berger, James O. (1976). "Admissible Minimax Estimation of a Multivariate Normal Mean with Arbitrary Quadratic Loss," Ann. Statist. 4, 223-226.

Bhattacharya, P.K. (1966). "Estimating the Mean of a Multivariate Normal Population with General Quadratic Loss Function," Ann. Math. Statist. 37, 1819-1824.

Bingham, C. and K. Larntz (1977). Comment. J. Amer. Statist. Assoc. 72, 97-102.

Bock, Mary Ellen (1975). "Minimax Estimators of the Mean of a Multivariate Normal Distribution," Ann. Statist. 3, 209-218.

Casella, George (1977). "Minimax Ridge Regression Estimation," Mimeograph Series #497, Dept. of Statistics, Purdue University.

Cohen, A. (1965). "Estimates of Linear Combinations of the Parameters in the Mean Vector of a Multivariate Distribution," Ann. Math. Statist. 36, 78-87.

Dempster, A.P., Martin Schatzoff and Nanny Wermuth (1977). "A Simulation Stidy of Alternatives to Ordinary Least Squares," J. Amer. Statist. Assoc. 72, 77-106.

Draper, N., and R.C. Van Nostrand (1979). "Ridge Regression and James-Stein Estimation: Review and Comments," Technometrics 21 (4), p. 451-466.

Efron, B., and C. Morris (1972). "Limiting Risk of Bayes and Empirical Bayes Estimators - Part II," J. Amer. Statist. Assoc. 67, 130-139.

____________________ (1976). "Families of Minimax Estimators of the Mean of a Multivariate Normal Distribution," Ann. Statist. 4, 11-21.

Hoerl, Arthur E., and Robert W. Kennard (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics 12, 55-67.

____________________ and Kent F. Baldwin (1975). "Ridge Regression: Some Simulations," Comm. in Statist. 4, 105-123.

James, W., and C. Stein (1961). "Estimation with Quadratic Loss," *Proceedings of the Fourth Berkeley Symposium, vol. I*, ed. by Jerzy Neyman, pp. 361-379. Berkeley and Los Angeles: University of California Press.

Lawless, J. F., and P. Wang (1976). "A Simulation Study of Ridge and Other Regression Estimators," *Comm. in Statist. A5*, 307-323.

McDonald, Gary C., and Diane I. Galarneau (1975). "A Monte Carlo Evaluation of Some Ridge Type Estimators," *J. Amer. Statist. Assoc. 70*, 407-416.

Stein, Charles (1956). "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution," *Proceedings of the Third Berkeley Symposium, Vol. I*, ed. by Jerzy Neyman, pp. 197-206. Berkeley and Los Angeles: University of California Press.

———————— (1973). "Estimation of the Mean of a Multivariate Normal Distribution," *Proceedings of the Prague Symposium on Asymptotic Statistics*, 345-381.

Stone, M. (1974). "Cross-validatory Choice and Assessment of Statistical Predictions," *J. Roy. Statist. Soc. B-36*, 111-147.

Strawderman, William E. (1978). "Minimax Adaptive Generalized Ridge Regression Estimators," *J. Amer. Statist. Assoc. 73*, 623-627.

Thisted, R., and C. Morris (1980). "Theoretical Results for Adaptive Ordinary Ridge Regression Estimators," University of Chicago Technical Report #94.

Footnotes

1. A possible refinement to using the predictor defined by $\hat{\beta}_S$ with $c = \hat{\sigma}^2(p-2)$ would employ the "limited translation" approach as discussed in Efron and Morris (1972, p. 136). Limiting the amount of shift for each coordinate of $\hat{\beta}_{OLS}$ toward the corresponding coordinate of $\hat{\beta}_S$ using a relevance function, ρ, leads to an estimator $\hat{\beta}_\rho$ and resulting predictor $X_0^T\hat{\beta}_\rho$. We also note that the estimator, $\hat{\beta}_S$, resulting from James and Stein (1961, p. 366) sets $c = \hat{\sigma}^2(p-2)(n-p+2)^{-1}(n-p)$. With this c (19) becomes $\sigma^2X_0^T(X^TX)^{-1}X_0\{1+(n-p+2)^{-1}(n-p)(p^2-4)\Gamma_1-2(n-p+2)^{-1}(n-2)(p-2)\Gamma_2\}$.

2. We recognize that these simplifying assumptions diminish the utility of the simulation study. In particular, certain estimators which differ in a more general model become equivalent. The study is only intended to be illustrative and suggestive. Certainly a more elaborate one might be undertaken. We also recognize that with these simplification expressions for the exact risks of <u>some</u> of the predictors below (ignoring restrictions) may be obtained using (17).

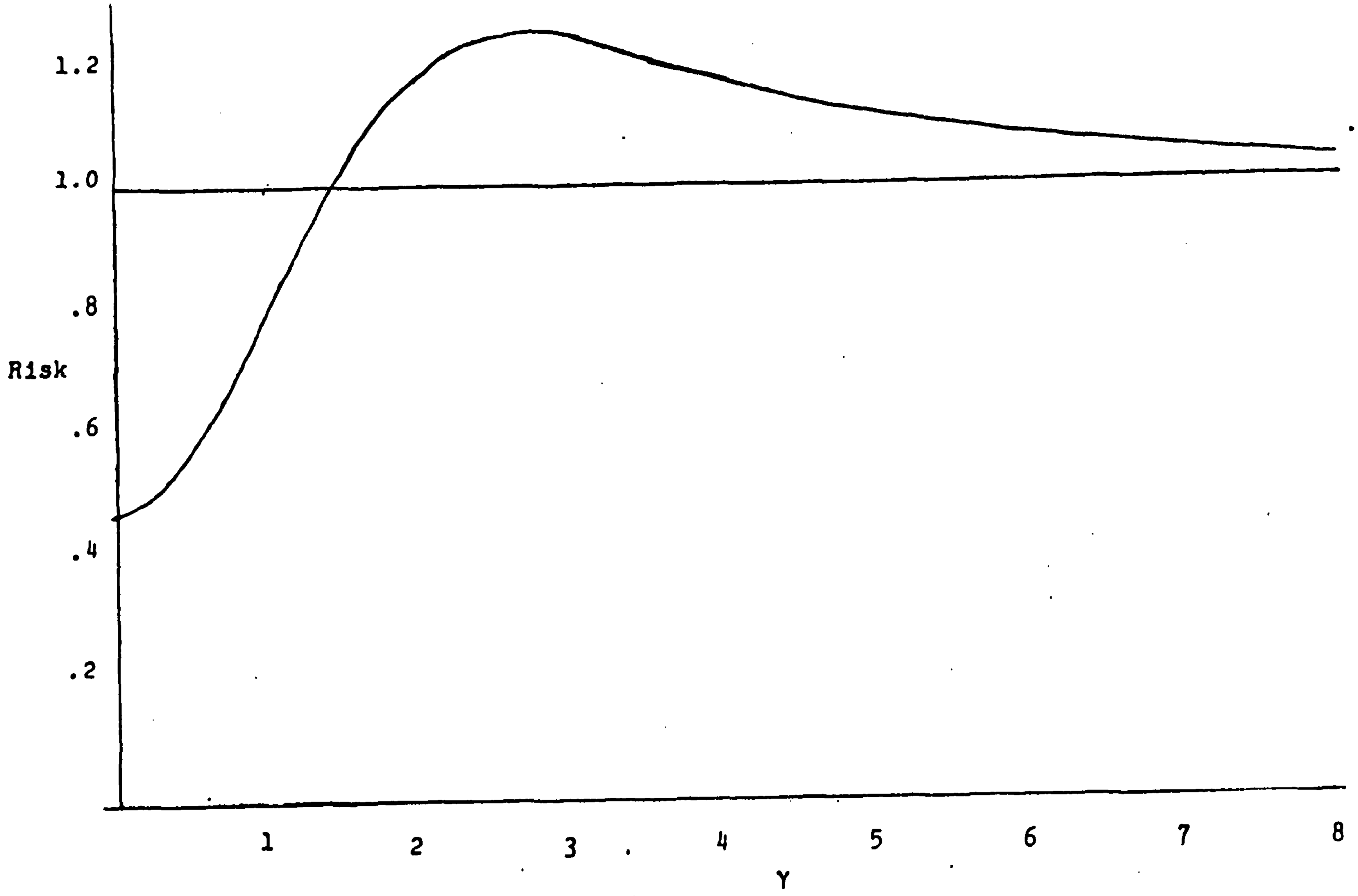


Figure 1: Risk of $(\hat{\gamma}^2+1)^{-1}\,\hat{\gamma}^3$

Table 1: A Brief Summary of Simulation Findings

		$\beta^T\beta$		
		small	moderate	large
p	3	$\hat{\beta}_{p-2}$ (95-99)*	$r > .45$, $\hat{\beta}_{HK}$ (40-60, $\downarrow$ in r) $r < .45$, $\hat{\beta}_{RM}$ (60-80, $\downarrow$ in r)	$r > .5$, $\hat{\beta}_{OLS}$ $.35 < r < .5$, $\hat{\beta}_{p-2}$ $.25 < r < .35$, $\hat{\beta}_{p-2}$, $\hat{\beta}_{MG}$, $\hat{\beta}_{HK}$** $.1 < r < .25$, $\hat{\beta}_{HK}$ $.01 < r < .1$, $\hat{\beta}_{RM}$ $r < .01$, $\hat{\beta}_{\hat{c}_0}$, $\hat{\beta}_{\hat{a}_0}$
	6	$\hat{\beta}_{p-2}$ (95-99)*	$r > .45$, $\hat{\beta}_{HK}$ (40-60, $\downarrow$ in r) $r > .45$, $\hat{\beta}_{RM}$ (60-80, $\downarrow$ in r)	$r > .4$, $\hat{\beta}_{OLS}$ $.25 < r < .4$, $\hat{\beta}_{p-2}$, $\hat{\beta}_{MG}$, $\hat{\beta}_{HK}$** $.1 < r < .25$, $\hat{\beta}_{HK}$ $.01 < r < .1$, $\hat{\beta}_{RM}$ $r < .01$, $\hat{\beta}_{\hat{c}_0}$, $\hat{\beta}_{\hat{a}_0}$
	10	$\hat{\beta}_{RM}$ (90-95)*	$r > .45$, $\hat{\beta}_{HK}$ (30-60, $\downarrow$ in r) $r < .45$, $\hat{\beta}_{RM}$ (60-90, $\downarrow$ in r)	$r > .4$, $\hat{\beta}_{OLS}$ $.25 < r < .4$, $\hat{\beta}_{p-2}$, $\hat{\beta}_{MG}$, $\hat{\beta}_{HK}$** $.15 < r < .25$, $\hat{\beta}_{HK}$ $.01 < r < .15$, $\hat{\beta}_{RM}$ $r < .01$, $\hat{\beta}_{\hat{c}_0}$, $\hat{\beta}_{\hat{a}_0}$

*Figures in parentheses indicate range of expected percent improvement over OLS predictor using "best" predictor.

**Are indistinguishable.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER 374	2. GOVT ACCESSION NO. AD-A168349	3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle) On The Use Of Ridge And Stein-Type Estimators In Prediction	5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Alan E. Gelfand	8. CONTRACT OR GRANT NUMBER(s) N00014-86-K-0156
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 1111	12. REPORT DATE May 21, 1986 13. NUMBER OF PAGES 24
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Ridge-type estimators, Stein-type estimators, Regression.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

PLEASE SEE FOLLOWING PAGE.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

TECHNICAL REPORT NO. 374

20. ABSTRACT

For the usual regression model with fixed regressors, there is a considerable literature devoted to alternatives to ordinary least squares estimators of the regression parameters. These alternatives are biased with "small" variances resulting in reduced mean square error over some (perhaps all) of the parameter space. Two prominent classes of such estimators are ridge-type and Stein-type estimators.

Consider the simplest prediction problem in this context, i.e. prediction at a single new vector of prediction values. ~~We calc~~ulate, the risk (squared error) for predictors based on estimators in the above families. While the ordinary least squares predictor is admissible, a simulation study reveals that over regions of the parameter space substantial reduction in risk is possible using estimators in these families. A simple preliminary procedure based upon the vector of prediction values is given to select a "good" estimator from these families. It is apparent that in multiple prediction a single choice of estimator need not be best.

END

DTIC

7—86